Beyond Tracking: Using Deep Learning to Discover Novel Interactions in Biological Swarms

Taeyeong Choi[1], Benjamin Pyenson[2], Juergen Liebig[2], and Theodore P. Pavlic[2,3,4]

[1] Lincoln Institute for Agri-food Technology, University of Lincoln, Riseholme Park, Lincoln, UK
[2] School of Life Sciences, Social Insect Research Group, Arizona State University, Tempe, AZ 85281, USA
[3] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85281, USA
[4] School of Sustainability, Arizona State University, Tempe, AZ 85281, USA
tchoi@lincoln.ac.uk, {bpyenson, jliebig, tpavlic}@asu.edu

Abstract. Most deep-learning frameworks for understanding biological swarms are designed to fit perceptive models of group behavior to individual-level data (e.g., spatial coordinates of identified features of individuals) that have been separately gathered from video observations. Despite considerable advances in automated tracking, these methods are still very expensive or unreliable when tracking large numbers of animals simultaneously. Moreover, this approach assumes that the human-chosen features include sufficient features to explain important patterns in collective behavior. To address these issues, we propose training deep network models to predict system-level states directly from generic graphical features from the entire view, which can be relatively inexpensive to gather in a completely automated fashion. Because the resulting predictive models are not based on human-understood predictors, we use explanatory modules (e.g., Grad-CAM) that combine information hidden in the latent variables of the deep-network model with the video data itself to communicate to a human observer which aspects of observed individual behaviors are most informative in predicting group behavior. This represents an example of augmented intelligence in behavioral ecology – knowledge co-creation in a human–AI team. As proof of concept, we utilize a 20-day video recording of a colony of over 50 *Harpegnathos saltator* ants to showcase that, without any individual annotations provided, a trained model can generate an "importance map" across the video frames to highlight regions of important behaviors, such as *dueling* (which the AI has no *a priori* knowledge of), that play a role in the resolution of reproductive-hierarchy re-formation. Based on the empirical results, we also discuss the potential use and current challenges to further develop the proposed framework as a tool to discover behaviors that have not yet been considered crucial to understand complex social dynamics within biological collectives.

Keywords: Deep Learning in Behavioral Ecology, Swarm Behavior, Explainable AI, Augmented Intelligence, Knowledge Co-creation

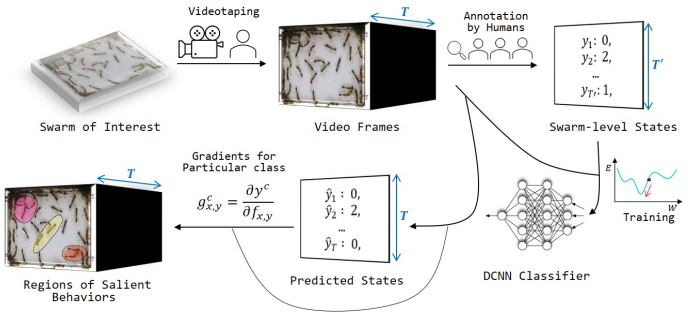


Fig. 1: Proposed usage of DCNNs, trained to predict global state of the swarm system from the entire view to later reveal key local observations by using the gradient between the learned local feature and the prediction output in the model.

1 Introduction

Deep Convolutional Neural Networks (DCNNs) have been widely adopted as the primary backbone of data-driven frameworks to solve complex problems in computer vision including object classification or detection and recognition of human actions [10], [14], [15]. The nature of their multi-layer structure has a powerful ability to automatically learn to identify key local features (e.g., edges) from raw pixels of images and combine into more meaningful concepts (e.g., pointy ears) to produce a final prediction output (e.g., dog), as the data is processed from the lowest layer through the higher ones [4]. In fact, this may imply that if the target data contains global information of biological swarms, lower-level visual properties such as locations, motions, and interactions of the entities could automatically be identified throughout the hierarchical layers during the training process. However, deep learning in behavioral biology has mostly been limited to building perceptive models to localize particular body parts of each entity to generate another input to a subsequent analysis model to capture motional concepts of individuals and perform a prediction for the entire swarm based on them [2], [5], [8], [11].

There can be two main challenges in this approach: 1) obtaining the individual feature labels can require a significant amount of human effort especially when a large group of system is examined, and 2) the choice of features relies heavily on prior knowledge of human experts in the biological system. To address these issues, as visualized in Fig. 1, we here suggest training the deep-network models to predict system-level states directly from generic graphical features from the entire view, which can be relatively inexpensive to gather, and examine the salient behavioral regularities discovered in the trained intermediate layers by using *gradient*-based explanation modules (e.g., Grad-CAM [13]). In other words, our proposal is to make more use of the aforementioned potential of DCNN to automatically discover fine-grained, individual-level motional

Fig. 2: Example of Grad-CAM in which the key regions are highlighted for class "Elephant"[1].

Fig. 3: Colony of 59 *H. saltator* as a testbed, with a foraging chamber accessed by the south tunnel.

patterns highly associated with macroscopic swarm properties so that the predictive model can later be queried about what these patterns are without being constrained by prior knowledge from human experts.

Specifically, in this paper, we propose the use of the explainable module Grad-CAM (Fig. 2) for biological research. Extending our previous work [3], we utilize a 20-day video recording of a colony of over 50 *Harpegnathos saltator* ants to demonstrate that without any individual annotations provided as input, the trained model can classify social stability of colonies while also generating an "importance map" across video frames to selectively highlight regions of interactions (e.g., *dueling*) as potentially important drivers of colony state.

2 Proposed Framework

Rather than training on small-scale features of individuals in videos, our approach trains a DCNN to predict coarse-grained, large-scale labels (y) from representations of generic features from video data. Any discrete, large-scale property can be used, such as whether a crowd [7] is about to riot. We use hierarchy state $y \in \{Stable, Unstable\}$ for a *H. saltator* colony [3]. Our n-layer classifier consists of m two-dimensional convolutional layers $\phi_{1 \leq \ell \leq m}$ followed by other types $\psi_{m+1 \leq \ell' \leq n}$, such as recurrent or fully-connected layers. Convolutional layers are used as feature extractors in this architecture since each output f_{ij} at ϕ_ℓ can compactly encode the local observation in a larger region ("receptive field") at previous layers $\phi_{\ell'' < \ell}$; i.e., a change in f_{ij} can imply the amplification or decrease of the motion pattern observed in the corresponding region.

For explanation of what visual regions are most important to the predictive model, Grad-CAM [13] is employed on K two-dimensional output feature maps, each denoted as $f^k \in \mathbb{R}^{h \times w}$, at a convolutional layer ϕ_ℓ to finally calculate the "importance map" M^c over the original input for a particular class c. In the technical aspect, ϕ_ℓ can be an arbitrary layer satisfying $\ell \in \{1, 2, ..., m\}$, but the layer ϕ_ℓ close to ϕ_m is typically chosen to access more abstract features with wider receptive fields than the ones available at lower layers $\phi_{\ell' < \ell}$. For brevity, we denote ϕ to be the chosen convolutional layer in the following descriptions.

To generate the importance map M^c, we first obtain the gradient g^c of the output y^c with respect to each feature map f^k from ϕ, i.e., $g^c_{ij} = \partial y^c / \partial f^k_{ij}$. Therefore, $g^c_{ij} > 0$ implies that enhancing the observational pattern encoded by f^k_{ij} increases the predicted likelihood of class c – the discovered pattern is "salient" for class c – and $g^c_{ij} \leq 0$ implies that the observation is considered irrelevant to the prediction of class c. Then, for each feature map f^k, Grad-CAM then uses this quantity to gain the averaged importance $a^c_k = (1/Z) \sum_i \sum_j g^c_{ij}$ (where Z is a normalization constant). Finally, the importance map M^c is computed by the weighted summation of feature maps:

$$M^c = \Gamma\left(\sum_k a^c_k \odot f^k \right) \tag{1}$$

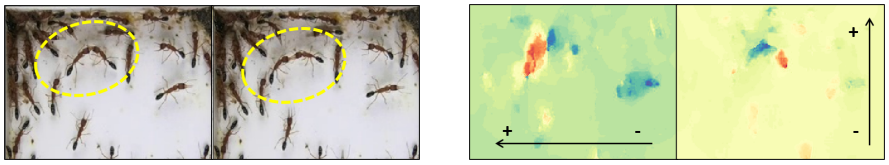
where $\odot$ is the element-wise multiplication, and $\Gamma(a) = a$ for $a > 0$ and $\Gamma(a) = 0$ otherwise. In Section 4, we also introduce a more restrictive Γ' that gates only the top 5% values so as to strictly verify whether key behaviors are effectively highlighted with the highest level of confidence. Also, M^c can be spatially upsampled to fit the original image of a desired size for visualization purpose.

3 Testbed Design with *H. saltator*

As in [3], a colony of *H. saltator* is utilized as a testbed to validate whether our proposed framework can reveal salient behavioral patterns. A conspicuous "unstable" state can be induced in this system through the removal of identified egg layers ("gamergates") [9] that triggers a hierarchy reformation process. During this process, aggressive interactions such as *dueling* [12] can be readily observed for several weeks until several mated workers activate their ovaries and start to lay eggs as the new gamergates, causing the colony to return to its nominal stable state [6]. We apply our framework to this system by building a binary-state classifier on the stability of the colony. We use the resulting deep-network model to identify important behaviors of interest and validate whether *dueling* (Fig. 4a) is discovered without *a priori* knowledge of it. Other behaviors identified by the system may then warrant further investigation by human researchers.

3.1 Video Data from Colonies Undergoing Stabilization

As shown in Fig. 3, each 20-day video was taken with an overhead camera to observe 59 *H. saltator* ants in plaster nests covered with glass. Due to a foraging chamber outside the view of the camera, not all ants are necessarily visible at all times, and some paralyzed crickets can be carried into the view. We artificially disturbed the reproductive hierarchy by removing all four preidentified gamergates after the second day of recording and further observed the process of hierarchy reformation until aggressive interactions almost disappeared in the last several days. Therefore, the video frames of the first 2 days are annotated with $y = Stable$, while the later ones of 18 days are all with $y = Unstable$.



(a) (b)

Fig. 4: (a) Example of two consecutive RGB frames cropped around a *dueling* interaction in yellow circle for visibility; (b) Horizontal and vertical optical flow vectors generated from (a), in each of which red (blue) are the regions of movement in the positive (negative) direction along the corresponding axis.

We follow the preprocessing method in [3] to extract from consecutive frames their optical flow, for which a pair of vectors encodes the horizontal and vertical transient movements from the input sequence (e.g., Fig. 4) [7]. Two optical flows in spatial resolution of 64×64 were computed every two minutes to use as an input x to the model, as each was obtained from two consecutive RGB frames 0.5 seconds apart in times. More details of the dataset are available online[5].

3.2 Deployed DCNNs with Grad-CAM

We use a classifier from our previous work [3] for the *one-class classification* task of predicting colony state. That colony-state classifier has an overall performance of 0.786 in the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) while only taking two consecutive optical flows as input. Moreover, colony-state predictions during the early period of first 6 days after the reproductive hierarchy is disturbed have higher AUC scores than 0.900 in average [3], indicating that the micro-scale graphical features identified by the deep network may be strong predictors of macro-scale state dynamics.

More specifically, the classifier we use has four 2D convolutional layers $\phi_{1:4}$ with 2D max pooling between consecutive layers, and six other types of layers $\psi_{5:10}$ follow to produce the estimated likelihood of unstable colony state. As described in Section 2, we then employ Grad-CAM on the feature maps from ϕ_4. For each generated importance map M^c, *bicubic* interpolation is applied to match the size of the frame image to overlay.

4 Results and Model Validation

As discussed in Section 3, we validate our approach by confirming that *dueling* behavior between ants is identified by the AI as strongly related to the unstable colony state. A model that can detect *dueling* with no prior knowledge of the behavior may identify other behavioral patterns that warrant further investigation.

[5] https://github.com/ctyeong/OpticalFlows_HsAnts

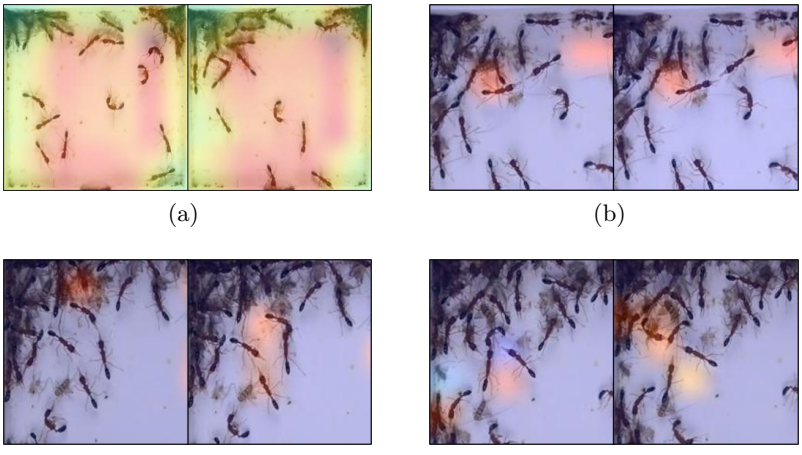
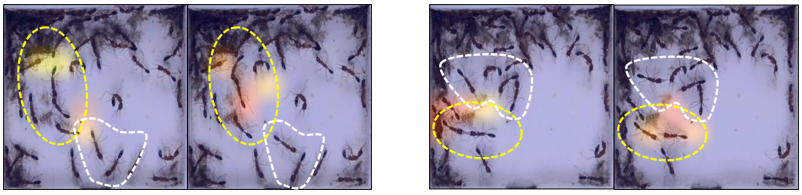


Fig. 5: (a): Heatmaps from Grad-CAM at two arbitrary times; (b), (c), (d): Three *dueling* examples captured by the top 5% impactful regions of red. Each pair shows two consecutive frames cropped around the interaction for clarity.

Figure 5a displays the heatmaps produced by the initial application of Grad-CAM with rectifier Γ. Grad-CAM identifies that the central area is more critical than the boundaries, and this general pattern is consistent over time despite changes in ant behaviors. This visualization indicates that, for the purpose of identifying changes in colony hierarchical state, the neural network has learned to ignore interactions near boundaries and instead focuses on interactions in the center of the area. Although this pattern matches intuition from human observations of these ants, it is too coarse to identify important behaviors.

We thus applied a filtered rectifier Γ' to only visualize regions of the top-5% positive gradients to identify the most dramatic responses in the generated heatmap to the ant motions, which resulted in more refined identifications of regions of importance. Figures 5b, 5c, and 5d show examples of *dueling* interactions detected by these highest gradients. Given that the deep network was not provided coordinates of the ants nor prior behavioral models of *dueling*, it is not surprising that the highlighted regions do not precisely identify specific ants in the interactions. Nevertheless, the network identifies general regions in close proximity to important behaviors. In particular, in Fig. 5c and 5d, more than two ants were engaged in *dueling*, but the detection region dynamically moved around them while they actively participated. These results support that the trained model has not overfit trivial attributes such as brightness or contrast of video but learned from ant behaviors themselves.

Figure 6a also shows the case where two duelers are captured as intended while other active ants who are simply showing swift turns nearby each other without direct interaction are ignored by our model. This indicates that the



(a) (b)

Fig. 6: Two examples in which *dueling* ants are detected (yellow dash line) while other active ones are ignored (white dash line).

DCNN classifier does not blindly take any type of movement into account for prediction; only relevant patterns are prioritized as features to utilize. Similarly, in Fig. 6b, two *dueling* ants are detected among a group of other non-dueling neighbors that are presenting rapid changes in motion and orientation. This example also demonstrates the ability of our trained model to filter out unimportant motion patterns even when a high degree of motion flow is present.

5 Summary, Discussion, & Future Work

We have proposed a deep-learning pipeline as a tool to uncover salient interactions among individuals in a swarm without requiring prior human knowledge about the behaviors or significant preprocessing effort devoted to individual tracking and behavioral coding. Our experimental results show that a trained classifier integrated with Grad-CAM can localize regions of key individual-scale interactions used by the classifier to make its colony-scale predictions. Validating our approach, identified behaviors, such as *dueling*, are the same behaviors that have been identified previously by human researchers without the aid of machine learning; however, our classifier discovered them without any prior guidance from humans. Thus, the library of other highlighted patterns from our pipeline can be used to generate new testable hypotheses of individual-to-colony emergence.

Our proposed approach greatly reduces human annotation effort as only macro-scale, swarm-level annotations are used in training. Significant effort is currently being used to develop machine-learning models for the subtask of tracking alone. Our approach suggests that tracking may, in many cases, be an unnecessary step that wastes both computational and human resources. Furthermore, our proposed approach reduces the risk of introducing human bias in the pre-processing of individual-level observations. Consequently, our example is a model of how human–AI observational teams can engage in knowledge co-creation – each providing complementary strengths and ultimately realizing the vision of augmented, as opposed to purely artificial, intelligence.

An important future direction is to further classify the highlighted patterns automatically discovered by these pipelines. Human behavioral ecologists can discriminate between peculiar interactions (e.g., *dueling*, *dominance biting*, and

policing [12]) that all may occur during the most unstable phases of reproductive hierarchy formation in *H. saltator* ants. Our method may have the ability to identify these behaviors, but it does not currently cluster similar identified patterns together and generate generalizable stereotypes that would be instructive to human observers hoping to identify these behaviors in their own future observations. Unsupervised learning methods could be adopted as a subsequent module to perform clustering and dimensionality reduction to better communicate common features of clusters, which may include patterns not yet appreciated by human researchers that are apparently useful in predicting swarm behavior.

References

[1] Grad-CAM class activation visualization (2020). URL https://keras.io/examples/vision/grad_cam/
[2] Bozek, K., Hebert, L., Mikheyev, A.S., Stephens, G.J.: Towards dense object tracking in a 2D honeybee hive. In: Proc. IEEE CVPR 2018 (2018)
[3] Choi, T., Pyenson, B., Liebig, J., Pavlic, T.P.: Identification of abnormal states in videos of ants undergoing social phase change. In: Proc. AAAI 2021 (2021)
[4] Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning. MIT press Cambridge (2016)
[5] Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D.: Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. eLife **8** (2019)
[6] Liebig, J., Peeters, C., Hölldobler, B.: Worker policing limits the number of reproductives in a ponerine ant. Proc. R. Soc. B **266**(1431) (1999)
[7] Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Proc. IEEE CVPR 2009 (2009)
[8] Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., Mathis, M.W.: Using deeplabcut for 3D markerless pose estimation across species and behaviors. Nat. Protoc. **14**(7) (2019)
[9] Peeters, C., Crewe, R.: Worker reproduction in the ponerine ant ophthalmopone berthoudi: an alternative form of eusocial organization. Behav. Ecol. Sociobiol. **18**(1) (1985)
[10] Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv:1804.02767 (2018)
[11] Romero-Ferrero, F., Bergomi, M.G., Hinz, R.C., Heras, F.J., de Polavieja, G.G.: Idtracker. ai: tracking all individuals in small or large collectives of unmarked animals. Nat. Methods **16**(2) (2019)
[12] Sasaki, T., Penick, C.A., Shaffer, Z., Haight, K.L., Pratt, S.C., Liebig, J.: A simple behavioral model predicts the emergence of complex animal hierarchies. Am. Nat. **187**(6) (2016)
[13] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proc. IEEE ICCV 2017, pp. 618–626 (2017)
[14] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
[15] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: Proc. ECCV 2016. Springer (2016)